Chengpo Yan

jesse-yan.github.io | cyan46@wisc.edu

Education

University of Wisconsin-Madison | Madison, WI
Ph.D. in Computer Science *Sep 2023 – May 2028*

University of Wisconsin-Madison | Madison, WI
B.S. in Computer Science & B.S. in Data Science *Sep 2019 – Dec 2022*

Research Interests

My research interests are in **Systems** and **Machine Learning**. My current focus is on optimizing the inference system of LLMs to improve throughput on cloud platforms and edge devices.

Publications

[**1**] BagPipe: Accelerating Deep Recommendation Model Training. Saurabh Agarwal, **Chengpo Yan**, Ziyi Zhang, Shivaram Venkataraman. [SOSP 2023]

[2] Deep-Learning-Based Segmentation of Keyhole in In-Situ X-ray Imaging of Laser Powder Bed Fusion. William Dong, Jason Lian, **Chengpo Yan**, Yiran Zhong, Sumanth Karnati, Qilin Guo, Lianyi Chen, Dane Morgan. [Materials 2024]

[3] Privacy-Aware Sharing of Raw Spatial Sensor Data for Cooperative Perception. Bangya Liu, Chenghao Jiang, **Chengpo Yan**, Suman Banerjee, Akarsh Prabhakara. [In submission]

Research Experience

Research Assistant with Prof. Suman Banerjee | Madison, WI
Deployment of Large Language Models on Tiny Devices *Sep 2024 – Present*

- Designed a token-driven, cross-accelerator inference system for vision-language models running on mobile SoCs.
- Optimized GPU-accelerated GEMM kernel on unified memory architecture, reducing the memory usage by 38%.
- Achieved a 2.3x throughput improvement by utilizing on-chip GPU and NPU.

Undergraduate Research Assistant with Prof. Shivaram Venkataraman | Madison, WI
BagPipe: Accelerating Deep Recommendation Model Training *May 2022 – Dec 2022*

- In this work, we systematically arrange and cache the embeddings required for a Deep Recommendation Model over distributed GPU clusters to optimize the time for fetching and updating embeddings given a parameter server training cluster.
- Reduced overall training time by 10% by overlapping the forward pass with NCCL-based cache synchronizations.
- Developed a policy for writing the embeddings evicted from cache on each worker to a parameter server under distributed GPU clusters, reducing the IO by 15%.
- Applied BagPipe to other Deep Recommendation Models, including W&D, D&C, DEEPFM, FGCNN, and CASER.
- Achieved up to 5x efficiency gain compared to the original DLRM proposed by Meta.

Undergraduate Research Assistant with Prof. Suman Banerjee | Madison, WI
Wi-Fi & Bluetooth Based Indoor Localization *Jul 2022 – Dec 2023*

- Designed a localization system that can use either RSSI or CSI, common indicators from Wi-Fi and Bluetooth devices.
- Tested different localization algorithms, including kNN, Stg, CSE, and Gaussian Kernel, achieving a median accuracy of two meters with the Gaussian Kernel approach.
- Used Apple's Airport to collect and generate RSSI dataset for the UW-Madison CS Department, which includes over a hundred access points.

Undergraduate Researcher with Prof. Dane Morgan @ Informatics Skunkworks | Madison, WI
Identifying Defect Formation Mechanisms from X-ray Imaging Data with Machine Learning *Mar 2022 – Dec 2022*

- Designed a continuous workflow of the entire identification and measurement stream, including image labeling and image measuring.
- Deployed BASNet, a boundary-aware segmentation network to label raw images for measurements, which shows promising accuracy of 91.24% for keyhole area and 92.81% for boundary shape.
- Implemented a parallel-based measuring tool for retrieving specific quantities given a labeled data.

Professional Experience

Teaching Assistant | Madison, WI
UW-Madison *Sept 2023 – Present*

- Spring 2024 to Fall 2025: CS354 Machine Organization and Programming.
- Fall 2023: CS540 Introduction to Artificial Intelligence.

Software Engineer Intern @ Skyworth | Shenzhen, China
Advisor: Dr. Chen *May 2021 – Aug 2021*

- Optimized Swaiot application, an EEUI based cross-platform application for controlling indoor devices through Skyworth AIOT system, achieved 9% improvement on loading time and 11% on response time
- Developed a theme switching functionality for Swaiot application, enabling the user to switch between light and dark mode.
- Adapted Swaiot application on various Skyworth devices, including Skyworth remote control, Skyworth tablet, and Skyworth table.

Web Development Club | Madison, WI
UW-Madison *Sep 2019 – May 2022*

- Contributed to the design of the UW-Madison CS Department's webpage, making it more user-friendly.
- Tutored web development frameworks and techniques.

Projects

Matrix Calculator
An Application for Matrix Calculation *Mar 2020 – May 2020*

- Implemented a JavaFX-based desktop application to provide accessible matrix calculation services for students.
- Planned the general workflow and designed the appearance of the application.

Mini-Twitter
A Mini Web Service that Emulates Twitter *May 2020 – Jun 2020*

- Implemented the web service with Ruby, Rails, MongoDB, Webpacker for online posting service.
- Developed Ajax-based frontend using ReactJS, Redux, and Saga to perform real-time notifications.

UniMatch
A Web Service that Offers Colleges Information and Provides Suggestions *Jun 2020 - Aug 2020*

- Implemented the web service with ReactJS, Webpacker, Flask, TensorFlow for matching up to five universities.
- Developed python-based backend with Flask and TensorFlow to find the most suitable university.
- Continuous deployment and integration with TravisCI.

Relevant Courses

CS: Algorithms, Artificial Intelligence, Computer Architecture, Data Analysis, Networks, Operating Systems, Human Computer Interaction, Virtual Reality
Math: Calculus, Discrete Math, Math in Data Science, Numerical Linear Algebra, Probability
Statistic: R in Data Science, Random Variables

Technical Skills

Programming: Java, Python, C, C#, C++, JavaScript, HTML/CSS, R
Frameworks: Spring, Hadoop, Spark, ReactJS, VueJS, NodeJS, Redux, MySQL, MongoDB, Docker, TensorFlow, PyTorch, CUDA, NCCL, TorchRec
Tools: Git, TravisCI, AWS, NSYS

Awards

- 2018 Meritorious Award on Mathematical Contest in Modeling
- Dean's List from 2019 to 2022 (6 semesters)